Transparency, Reproducibility, and Replicability in Federal Statistics: The Case of LEHD

Robert T. Sienkiewicz, Ph.D., MBA
Assistant Center Chief
Longitudinal Employer-Household Dynamics Program
Center for Economic Studies

Acknowledgements and Disclaimer

I thank John Abowd, John Eltinge, Lucia Foster, Matthew Graham, Erika McEntarfer, Camille Norwood, and Lars Vilhuber for providing me with valuable insights and direction on transparency, reproducibility and replicability in federal statistical organizations.

The views expressed here are those of the author and do not necessarily represent the policies of the United States Census Bureau.

Qualitative Description of Transparency, Reproducibility and Replicability

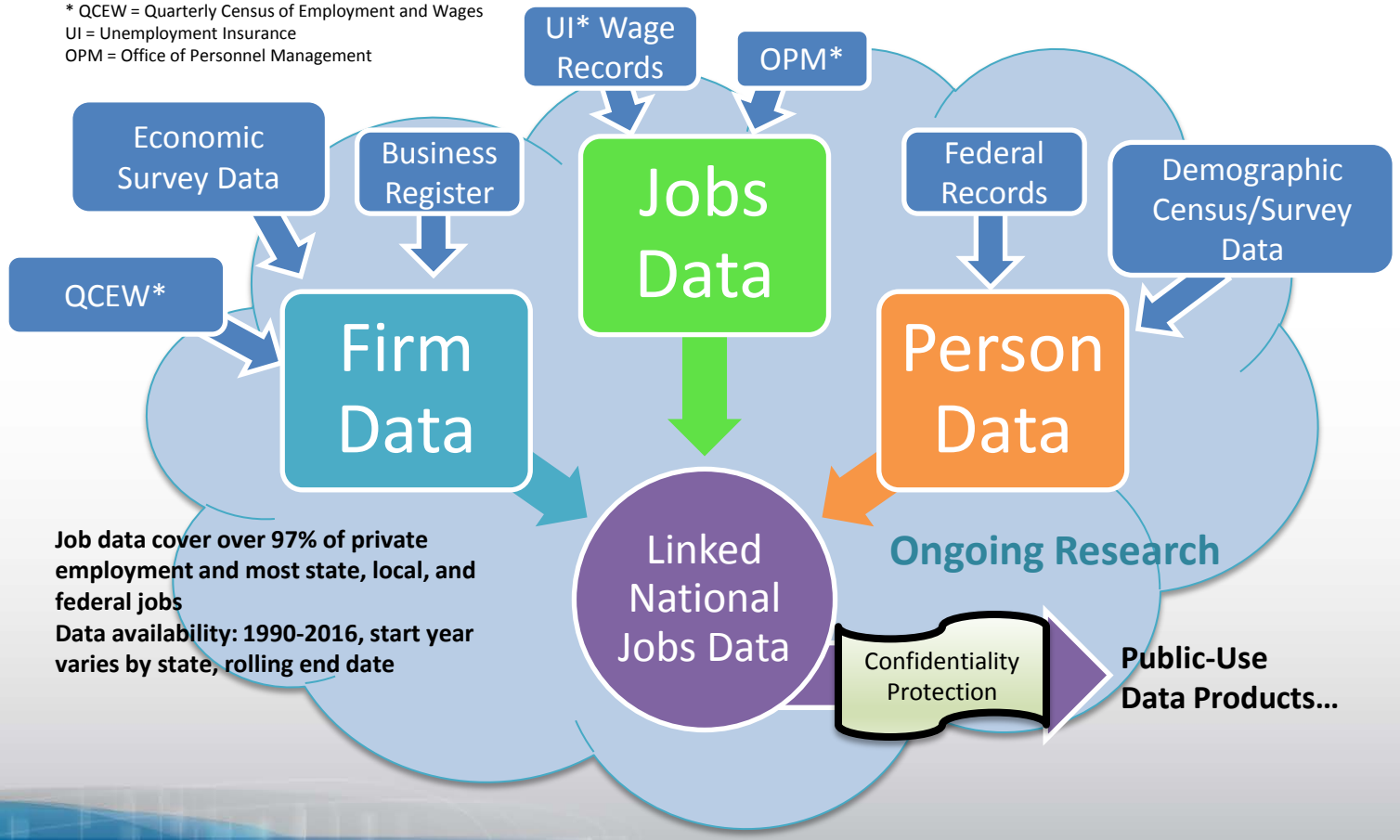
- Transparency
  - Is the data provided to the public in a comprehensible, accessible, and timely manner?
- Reproducibility
  - Do we make a recipes available that allow for the same results?
- Replicability
  - Can we replicate the results?

Mission and Structure

LEHD (Longitudinal Employer-Household Dynamics) Program

- Connects administrative records with census and survey data to produce ***new*** public-use data products as well as microdata for research.
- Links information on employers with employees to provide new data on the US economy. LEHD constructs unique linked employer-employee data for the United States and public-use products from this data (QWI, OTM, OTM-EM, J2J)
- Its unique employer-employee linked dataset covers 97% of jobs in the United States.

LEHD Data Infrastructure



Transparency and Reproducibility

- Transparency
  - Code
  - Description of methodology at different levels
  - Providing a comprehensive suite of documents
- Reproducibility
  - SOPs
  - Availability of Raw Inputs (UI Wage, ES202)
  - System Architecture
  - Code Versioning
  - Redundancy

Replicability: Significance

- Ability to generate, on demand, released data, recreated from original data inputs
- Fostered efficiency in the production of statistical data products
- Stood test of time
- Maintained Census Bureau's reputation as reliable provider of data.

Importance of Metadata System

- Documents and curates all data inputs, outputs, and processes
- Promote reliability, adaptability and replicability
- Statistical, IT and Computational Issues

Metadata Example

Traceability of Files Contributing to a Published QWI Release: LEHD Metadata Example

State: Arkansas (AR) Release: R2017Q1 Demographic: SA_F (Sex by Age for All Firm Types)

Each data release is accompanied by a file specifying a compact notation for metadata. For instance, the R2017Q1 release of Arkansas QWI by sex and age for all firm types would have a file called https://lehd.ces.census.gov/pub/ar/R2017Q1/DVD-sa_f/version_sa_f.txt with the following content:

QWISA_F AR 05 2002:3-2016:2 V4.1.1 R2017Q1 qwipu_ar_20170223_1545

Description of Version File Components		
#	Component	Description
1	QWISA_F	Series Type
2	AR	State Postal Abbreviation
3	05	State FIPS
4	2002:3-2016:2	Data Date Range
5	V4.1.1	Schema Version (see https://lehd.ces.census.gov/data/schema/)
6	R2017Q1	Release Quarter
7	qwipu_ar_20170223_1545	Unique Vintage Identifier (process_id) (state) (vintage)

How Used?

- Tracing back data problems
  - Provides scope of affected vintages
- Core of snapshot creation
  - Input into replicable research

Lessons Learned/Lessons Shared

- Technical
  - Developing a system that is “tight” but “flexible”
  - Computational Challenges
- Human and Organizational Issues
  - Developing a team with the right skill sets
  - Interface between system architect and programming team

Next Steps/Looking for Insights

- Technical
  - Tracking intermediate data file size
  - Automated scheduling
- Human and Organizational
  - Workflow Optimization
  - Legacy thinking
- Expansion to other LEHD data products
  - LODES, J2J

Thank You

Contact information:

robert.sienkiewicz@census.gov

(301) 763-1234